Siya Das

Assignment Day 7 – HR Analytics Case

Launching

```
import pandas as pd
```

```
import numpy as np
import matplotlib.pyplot as plt
```

```
import seaborn as sns
```

```
from plotly import __version__
print(__version__)
import cufflinks as cf
from plotly.offline import download_plotlyjs,init_notebook_mode,plot,iplot
init_notebook_mode(connected=True)
cf.go_offline()
```

```
4.8.2
```

```
import scipy.stats as stats
```

Loading the csv file (dataset).

```
df=pd.read_csv(r"C:\Users\siyad\AppData\Local\Temp\Temp2_Day-7-20200715T141046Z-001.zip\Day-7\Assignment\general_data.csv")
```

```
df.head()
```

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	NumCompaniesWo
0	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	1	Female	...	
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sciences	1	2	Female	...	
2	32	No	Travel_Frequently	Research & Development	17	4	Other	1	3	Male	...	
3	38	No	Non-Travel	Research & Development	2	5	Life Sciences	1	4	Male	...	
4	32	No	Travel_Rarely	Research & Development	10	1	Medical	1	5	Male	...	

5 rows × 24 columns

```
df.columns
```

```
Index(['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',
       'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',
       'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',
       'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager'],
      dtype='object')
```

Data Cleaning and pre-processing

- Checking for null values
- Dropping the null values

```
df.isnull().sum()
```

```
Age                       0
Attrition                 0
BusinessTravel            0
Department                0
DistanceFromHome          0
Education                 0
EducationField            0
EmployeeCount             0
EmployeeID                0
Gender                    0
JobLevel                  0
JobRole                   0
MaritalStatus             0
MonthlyIncome             0
NumCompaniesWorked       19
Over18                    0
PercentSalaryHike         0
StandardHours             0
StockOptionLevel          0
TotalWorkingYears         9
TrainingTimesLastYear     0
YearsAtCompany            0
YearsSinceLastPromotion   0
YearsWithCurrManager      0
dtype: int64
```

```
df1=df.dropna()
```

```
df1.isnull().sum()
```

```
Age                       0
Attrition                 0
BusinessTravel            0
Department                0
DistanceFromHome          0
Education                 0
EducationField            0
EmployeeCount             0
EmployeeID                0
Gender                    0
JobLevel                  0
JobRole                   0
MaritalStatus             0
MonthlyIncome             0
NumCompaniesWorked        0
Over18                    0
PercentSalaryHike         0
StandardHours             0
StockOptionLevel          0
TotalWorkingYears         0
TrainingTimesLastYear     0
YearsAtCompany            0
YearsSinceLastPromotion   0
YearsWithCurrManager      0
dtype: int64
```

- Checking for duplicates and dropping, if any. In this dataset, no duplicates were found.

```
df1.duplicated().sum()
```

```
0
```

Univariate Analysis

- **Of all employees**

Getting the count, mean, standard deviation, min, quartiles and max values by describing the numerical variables in the dataset.

```
df3=df1[['Age','DistanceFromHome','Education','MonthlyIncome','NumCompaniesWorked','PercentSalaryHike','TotalWorkingYears',
         'TrainingTimesLastYear','YearsAtCompany','YearsSinceLastPromotion','YearsWithCurrManager']]
```

```
df3.describe()
```

	Age	DistanceFromHome	Education	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	TrainingTimesLastYear
count	4382.000000	4382.000000	4382.000000	4382.000000	4382.000000	4382.000000	4382.000000	4382.000000
mean	36.933364	9.198996	2.912369	65061.702419	2.693291	15.210634	11.290278	2.798266
std	9.137272	8.105396	1.024728	47142.310175	2.497832	3.663007	7.785717	1.289402
min	18.000000	1.000000	1.000000	10090.000000	0.000000	11.000000	0.000000	0.000000
25%	30.000000	2.000000	2.000000	29110.000000	1.000000	12.000000	6.000000	2.000000
50%	36.000000	7.000000	3.000000	49190.000000	2.000000	14.000000	10.000000	3.000000
75%	43.000000	14.000000	4.000000	83790.000000	4.000000	18.000000	15.000000	3.000000
max	60.000000	29.000000	5.000000	199990.000000	9.000000	25.000000	40.000000	6.000000

- The mean age of the employees was 37 and IQR 13.
- Mean monthly income was 65061 and IQR 54000 which implies that the employees incomes ranged between a large number of values and therefore, large number of positions.

```
df3.median()
```

```
Age                        36.0
DistanceFromHome            7.0
Education                   3.0
MonthlyIncome           49190.0
NumCompaniesWorked          2.0
PercentSalaryHike          14.0
TotalWorkingYears          10.0
TrainingTimesLastYear       3.0
YearsAtCompany              5.0
YearsSinceLastPromotion     1.0
YearsWithCurrManager        3.0
dtype: float64
```

```
df3.mode()
```

	Age	DistanceFromHome	Education	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany
0	35	2	3	23420	1.0	11	10.0	2	5

```
df3.var()
```

```
Age                        8.348974e+01
DistanceFromHome           6.569744e+01
Education                  1.050068e+00
MonthlyIncome              2.222397e+09
NumCompaniesWorked         6.239165e+00
PercentSalaryHike          1.341762e+01
TotalWorkingYears          6.061739e+01
TrainingTimesLastYear      1.662558e+00
YearsAtCompany             3.756894e+01
YearsSinceLastPromotion    1.040059e+01
YearsWithCurrManager       1.274257e+01
```

```
df3.skew()
```

```
Age                        0.413048
DistanceFromHome           0.955517
Education                 -0.288977
MonthlyIncome              1.367457
NumCompaniesWorked         1.029174
PercentSalaryHike          0.819510
TotalWorkingYears          1.115419
TrainingTimesLastYear      0.551818
YearsAtCompany             1.764619
YearsSinceLastPromotion    1.980992
YearsWithCurrManager       0.834277
dtype: float64
```

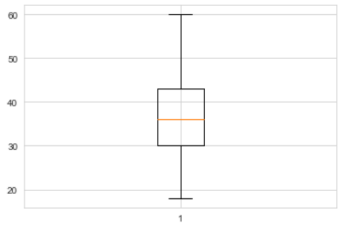
```
df3.kurt()
```

```
Age                       -0.409517
DistanceFromHome          -0.230691
Education                 -0.565008
MonthlyIncome              0.990836
NumCompaniesWorked         0.014307
PercentSalaryHike         -0.306951
TotalWorkingYears          0.909316
TrainingTimesLastYear      0.494215
YearsAtCompany             3.930726
YearsSinceLastPromotion    3.592162
YearsWithCurrManager       0.170703
dtype: float64
```

All were right skewed except age, which was left skewed.

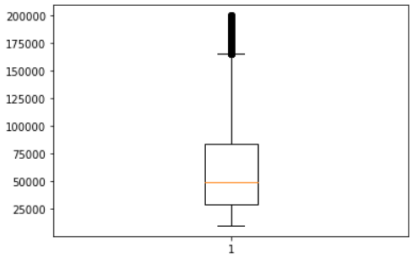
All variables were leptokurtic except Age, DistanceFromHome, Education and PercentSalaryHike, which were platykurtic.

```
plt.boxplot(df1['Age'])
```



There were no outliers with normal distribution.

```
plt.boxplot(df1['MonthlyIncome'])
```




```
stats.iqr(df1['MonthlyIncome'])
```

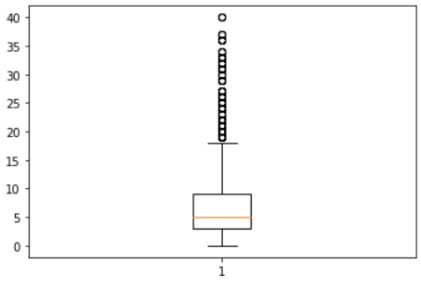
```
54680.0
```

MonthlyIncome had few outliers.

It is right skewed.

IQR is 54680.

```
plt.boxplot(df1['YearsAtCompany'])
```




```
stats.iqr(df1['YearsAtCompany'])
```

```
6.0
```

YearsAtCompany had outliers.

It is right skewed with IQR 6.

- **Of employees with Attrition Yes**

the count, mean, standard deviation, min, quartiles and max values was obtained by describe function after which the following was done.

```
df1['Attrition'].value_counts()
```

```
No     3677
Yes     705
Name: Attrition, dtype: int64
```

```
atr_yes=df1[df1['Attrition']=='Yes']
```

```
atr_yes['Gender'].value_counts()
```

```
Male      437
Female    268
Name: Gender, dtype: int64
```

```
atr_yes['JobRole'].value_counts()
```

```
Sales Executive              165
Research Scientist           158
Laboratory Technician        125
Healthcare Representative     56
Research Director             54
Manufacturing Director        48
Manager                       42
Sales Representative          36
Human Resources               21
Name: JobRole, dtype: int64
```

```
atr_yes['JobLevel'].value_counts()
```

```
2    283
1    250
3     96
4     51
5     25
Name: JobLevel, dtype: int64
```

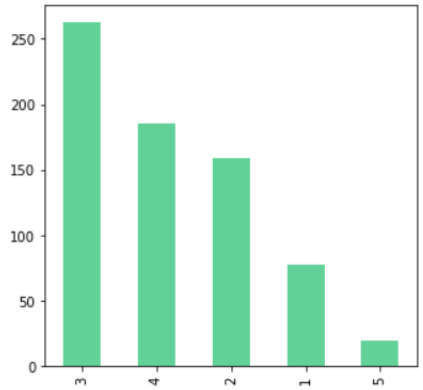
437 employees with Attrition Yes were male.

The two Job Roles where Attrition was seen most were Sales Executive and Research Scientist.

The two Job Levels where attrition was seen most were 1 and 2.

```
atr_yes['Education'].value_counts().plot(kind="bar", figsize=(5,5), color="#61d199")
```

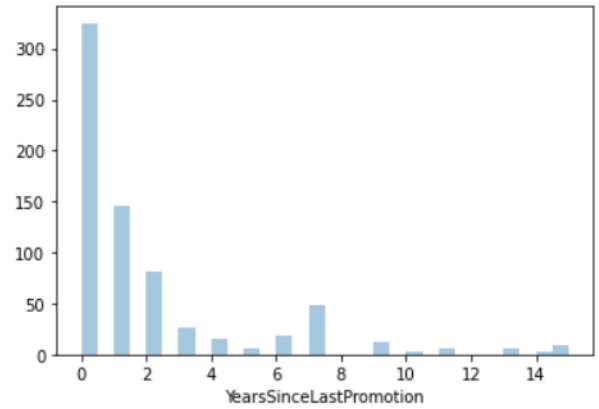
```
<matplotlib.axes._subplots.AxesSubplot at 0x2a45fbe9358>
```



The Attrition was highest in Education level 3 i.e. Bachelor.

```
sns.distplot(atr_yes['YearsSinceLastPromotion'],kde=False,bins=30)
```

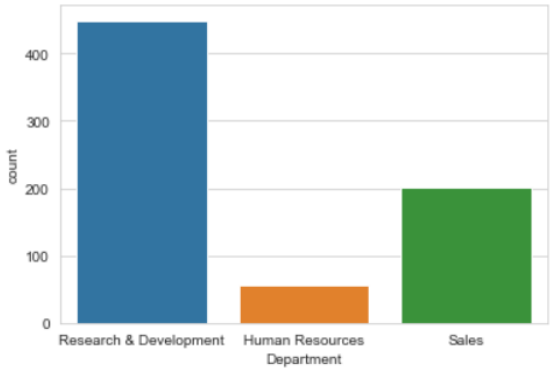
```
<matplotlib.axes._subplots.AxesSubplot at 0x2a45fed4780>
```



Most employees with attrition yes and less than 5 years since their last promotion.

```
sns.set_style('whitegrid')
sns.countplot(x='Department',data=atr_yes)
```

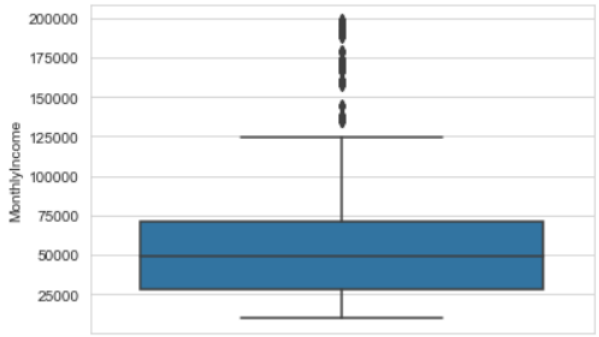
```
<matplotlib.axes._subplots.AxesSubplot at 0x2a460067a90>
```



Maximum Employees with Attrition Yes were in Research and Development, followed by Sales.

```
sns.boxplot(y='MonthlyIncome',data=atr_yes)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a462c57240>
```




```
stats.iqr(atr_yes['MonthlyIncome'])
```

```
42600.0
```

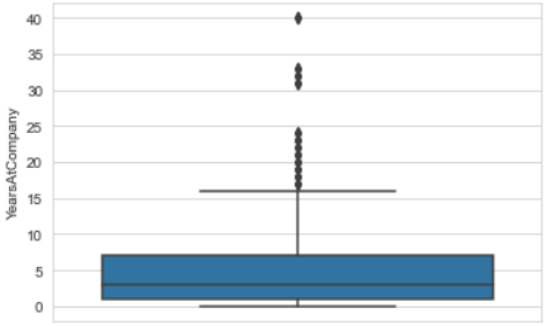
There were outliers.

Monthly income is right skewed.

IQR is 42,600, which implies that there was not much difference in the range of monthly income wrt Attrition or no attrition.

```
sns.boxplot(y='YearsAtCompany',data=atr_yes)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a46225a7b8>
```




```
stats.iqr(atr_yes['YearsAtCompany'])
```

```
6.0
```

There were outliers.

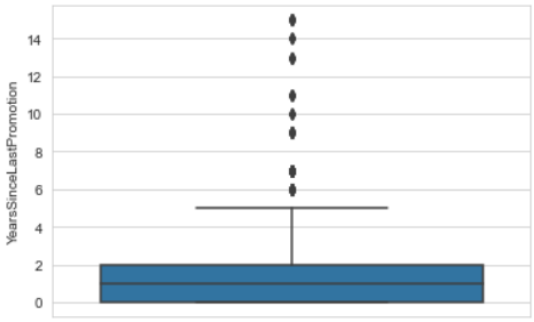
Years at Company is right skewed.

IQR is 6 but there are many outliers above 15 years.

Most employees left within 6 years with few who left after 15 or more years.

```
sns.boxplot(y='YearsSinceLastPromotion',data=atr_yes)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a4633129b0>
```



There were outliers.

IQR is 2 and the mean is also 2.

Most employees who left had not got promotion in the last 2 years.

```
stats.iqr(atr_yes['YearsSinceLastPromotion'])
```

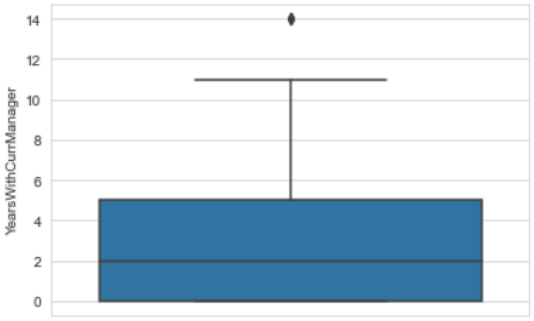
```
2.0
```

```
atr_yes['YearsSinceLastPromotion'].mean()
```

```
1.9602836879432624
```

```
sns.boxplot(y='YearsWithCurrManager',data=atr_yes)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a4637410f0>
```



There were very few outliers.

YearsWithCurrentManager is right skewed.

IQR is 5 and mean is 3.

Most employees were under the current manager for around 3 years, but range was between 0-5 years.

```
stats.iqr(atr_yes['YearsWithCurrManager'])
```

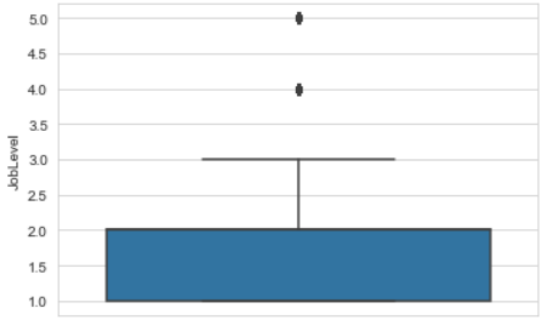
```
5.0
```

```
atr_yes['YearsWithCurrManager'].mean()
```

```
2.8652482269503547
```

```
sns.boxplot(y='JobLevel',data=atr_yes)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a4632c3208>
```



There were very few outliers.

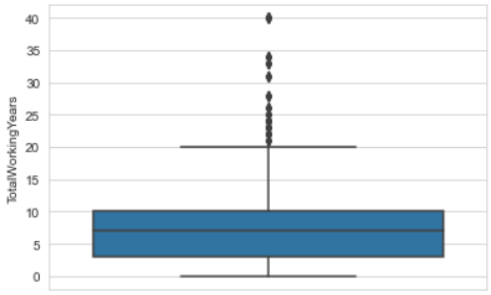
IQR is 1 which means most employees had job level either 1 or 2 with very few more than 3.

```
stats.iqr(atr_yes['JobLevel'])
```

```
1.0
```

```
sns.boxplot(y='TotalWorkingYears',data=atr_yes)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a463d6c1d0>
```




```
stats.iqr(atr_yes['TotalWorkingYears'])
```

```
7.0
```

```
atr_yes['TotalWorkingYears'].mean()
```

```
8.273758865248228
```

There were many outliers.

It is left skewed.

IQR is 7 and mean is 8.3.

Most employees who left were in the range of 3-10 years in the company.

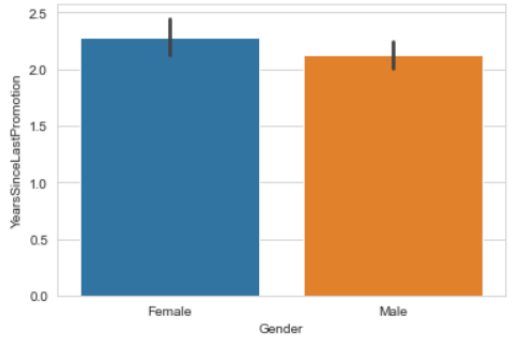
Mean being 8.2 implies that there were many outliers with more than 20 years in the company.

Bivariate Analysis

- **Of all employees**

```
sns.barplot(x='Gender',y='YearsSinceLastPromotion',data=df1,estimator=np.mean)
```

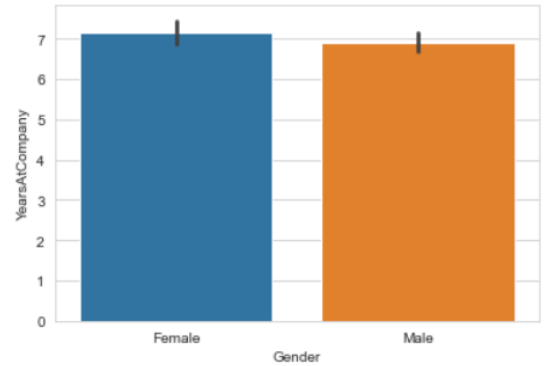
```
<matplotlib.axes._subplots.AxesSubplot at 0x2a4652a4358>
```



There was almost no difference between male and female employees wrt Years since last promotion.

```
sns.barplot(x='Gender',y='YearsAtCompany',data=df1,estimator=np.mean)
```

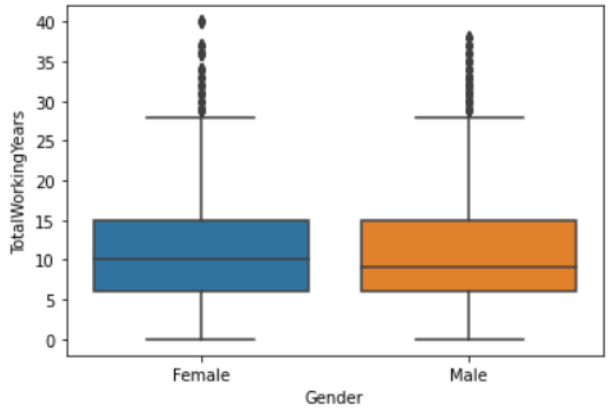
```
<matplotlib.axes._subplots.AxesSubplot at 0x2a4634b8128>
```



There was almost no difference between male and female employees wrt Years at company.

```
sns.boxplot(x='Gender',y='TotalWorkingYears',data=df1)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a45dbfc9e8>
```



There was almost no difference between male and female employees wrt Total working years.

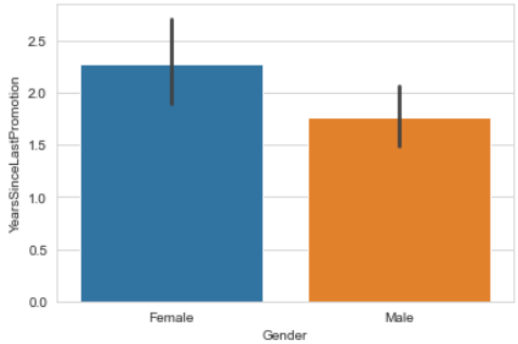
Male was slightly more right skewed.

- Scatterplots showed no regression between any 2 variables in the datset.

- **Of employees with Attrition Yes.**

```
sns.barplot(x='Gender',y='YearsSinceLastPromotion',data=atr_yes,estimator=np.mean)
```

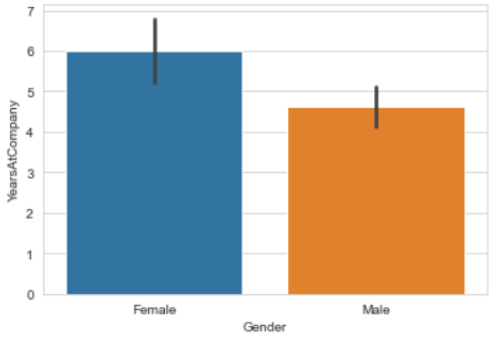
```
<matplotlib.axes._subplots.AxesSubplot at 0x2a4632a97b8>
```



There was considerable difference seen in gender wrt YearsSince lastPromotion where the mean years of female was more than 2 and male being less than 1.5.

```
sns.barplot(x='Gender',y='YearsAtCompany',data=atr_yes,estimator=np.mean)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a4630d8f98>
```

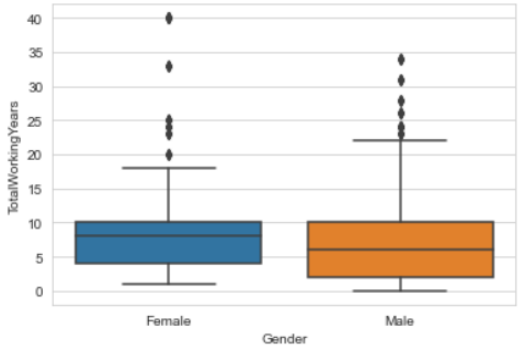


There was considerable difference seen in gender wrt YearsAtCompany where the mean years of female was 6 and male being less than 4.5.

Siya Das

```
sns.boxplot(x='Gender',y='TotalWorkingYears',data=atr_yes)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a4635e6b70>
```



There was a little difference seen in gender wrt Total working years.

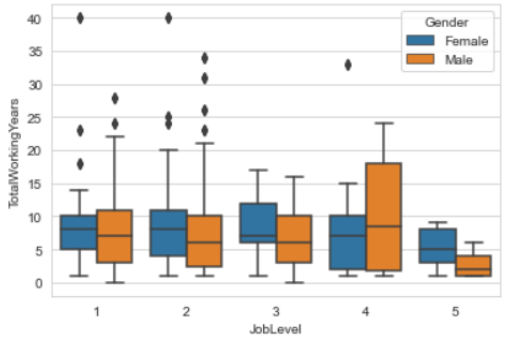
Female was left skewed while male was almost symmetrical.

Outliers were dispersing more in female.

IQR higher in male.

```
sns.boxplot(x='JobLevel',y='TotalWorkingYears',data=atr_yes,hue='Gender')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a46329cb00>
```



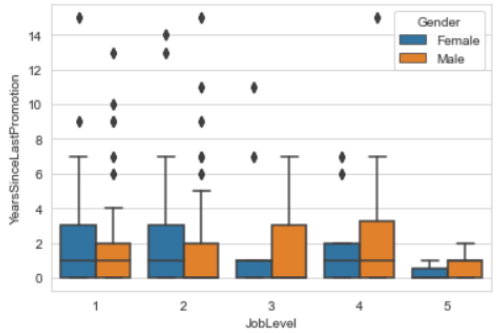
The boxplot of total working years of gender was plotted with job level in the x axis.

In job level 1, IQR of female was more than male while in 2 and 3, there wasn't much difference.

In level 4, the IQR of male was much higher while in level 5, it was much lower.

```
sns.boxplot(x='JobLevel',y='YearsSinceLastPromotion',data=atr_yes,hue='Gender')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a4631ed780>
```



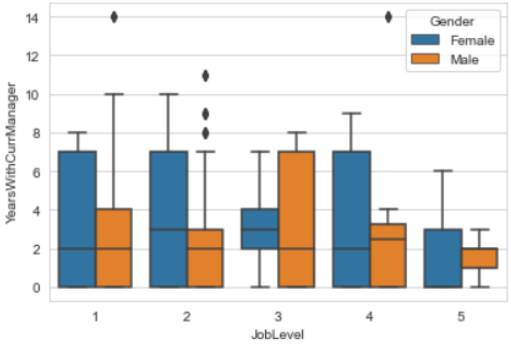
The boxplot of years since last promotion of gender was plotted with job level in the x axis.

In job levels 1 and 2, IQR was more in females with outliers being more in males.

In levels 3, 4 and 5, IQR was more in male.

```
sns.boxplot(x='JobLevel',y='YearsWithCurrManager',data=atr_yes,hue='Gender')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a4656a7898>
```



The boxplot of years with current manager of gender was plotted with job level in the x axis.

In all the job levels except 3, IQR of female was much more than that of their counterpart, the opposite of this was observed in Job level 3.

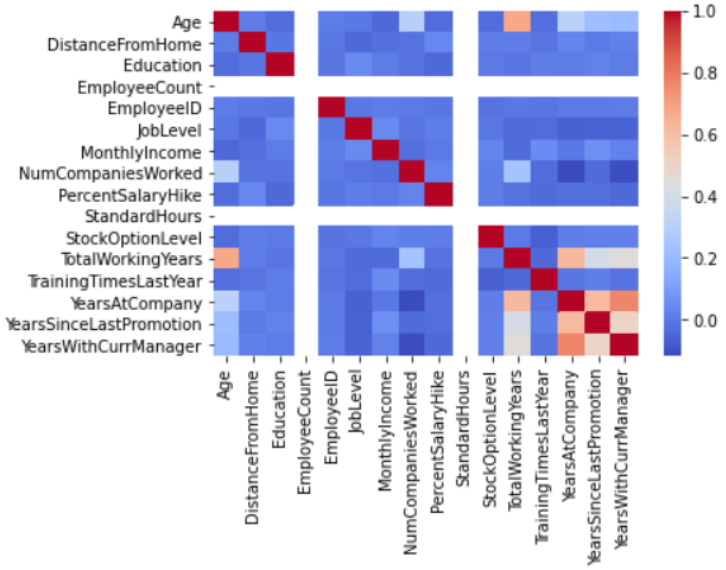
Siya Das

Multivariate Analysis

- **Of all employees**

```
dc=df1.corr()
sns.heatmap(dc,cmap='coolwarm')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a435f0c710>
```



The correlation matrix was drawn.

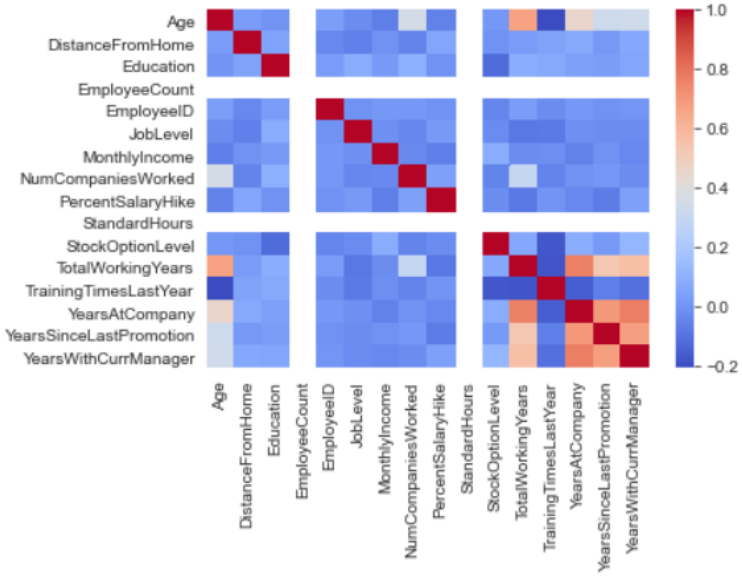
Correlation was observed in the lower right region mostly i.e YearsAtCompany, YearsSinceLastPromotion and YearsWithCurrentManager.

Employees were there for more years in the company and having more years since last promotion.

- **Of employees with Attrition Yes**

```
sns.heatmap(atr_yes.corr(),cmap='coolwarm')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a4652a46d8>
```



The correlation matrix was drawn.

Correlation was observed in the lower right region mostly i.e YearsAtCompany, YearsSinceLastPromotion and YearsWithCurrentManager.

There was more correlation of the above mentioned variables with Total working years as well.

The results were more or less like the correlation matrix obtained of all employees.